CARDIOVASCULAR RISK PREDICTION

by
Shaik Ahmad Basha
Data science trainee,
AlmaBetter, Bangalore.

Abstract:

The Framingham Heart Study is a long-term, ongoing cardiovascular cohort study of residents of the city of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants. Prior to the study almost nothing was known about the epidemiology of hypertensive or arteriosclerotic cardiovascular disease. Much of the now-common knowledge concerning heart disease, such as the effects of diet, exercise, and common medications such as aspirin, is based on this longitudinal study. It is a project of the National Heart, Lung, and Blood Institute, in collaboration with (since 1971) Boston University. Various health professionals from the hospitals and universities of Greater Boston staff the project.

.

Problem Statement

The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. ▪ The classification goal is to predict whether thepatient has a 10-year risk of future coronary heart disease (CHD). ▪ The dataset provides the patients' information. It includes over 4000 records and 17 attributes.

Data Description

The dataset contains patient's information and target feature.

The dataset has 3390 rows and 17columns. Those Features are

- Sex – Male or female
- Age- Age of the patient
- Is_smoking – whether the patient smokes or not
- cigsPerDay- how many cigarettes does the patient smokes
- BPMeds – whether the patient was on bp medications
- prevalentStroke – whether the patient had stroke previously
- prevalentHyp – whether the patient was hypertensive
- diabetes – whether the patient is diabetic or not
- totChol- Cholestrol level of patient
- sysBP – Systolic blood pressure
- diaBP – Diastolic blood pressure
- BMI – Body mass index
- heartRate - Hearrrate
- glucose- glucose level of patient
- TenYearCHD – whether the person has Ten year Coronary heart disease risk or not

Introduction:

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, anestimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three quarters of these deaths took place in low- and middle-income countries. Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the mostfatal. In the United States, for example, it is estimated that someone has a heart attack every 40 seconds and about 805,000 Americans have a heart attack every year.

Doctors and scientists alike have turned to machine learning (ML) techniques to develop screening tools and this is because of their superiority in pattern recognition and classification as compared to other traditional statistical approaches. In this project, We will be giving you a walk through on the development of a screening tool for predicting whether a patient has a 10-year risk of developing coronary heart disease (CHD) using different Machine Learning techniques.

Null Values Treatment

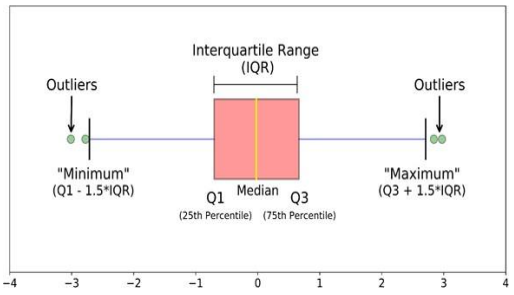
Missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data area common occurrence and can have a significant effect on the conclusions that can be drawn from the data. Continuing to that we found missing observation in sevencolumns which we further treated with its median value that corresponds to that column.

Exploratory Data Analysis

The exploratory data analysis that we performed on our train dataset helped us to realize how different features in our dataset influence the target variable.

Removing Outliers

A data point that varies greatly from other results is referred to as an outlier**.** An outlier may also be described as an observation in our data that is incorrect or abnormal as compared to other observations. To find outliers, we can simply plot the box plot for every feature present in the datasets. Outliers are pointsthat are outside of the minimum and maximum values. We have treated the outlier that lies away from the upper boundary and lower boundary with its median values



Data Cleaning and Manipulation

Manipulation of data is the process of manipulating or changing information to make it more organized and readable. Checking and dropping the duplicates values present in the datasets. Look over to some categorical features along with applying value counts to give us some intuition about that column. Defining a label encoding in two columns refers to converting the labels into numeric form so as to convert it into the machine-readable form.

Univariate Analysis

Univariate analysis is an analysis used on one variable with the aim of finding out and identifying the characteristics of the variable. This analysis is the most basic analysis technique that is often used in various types of research. We can see whether the data we use at a glance is normally distributed, left-winged, right-handed, there are outliers, etc. Knowing the size of concentration, size of distribution, and mean-median from a data set.

Bivariate Analysis

To implement bi-variate analysis using python. Bi-variate Analysis finds out the relationship between independent variables with single dependent variables. Using bi-variate analysis association and dissociation between variables at a pre-defined significance level. We can perform bi-variate analysis for any combination of categorical and continuous variables.

Removing Multicollinearity

- Multicollinearity is a statistical concept where several independent variables in a model are correlated.
- Two variables are considered to be perfectly collinear if their correlation coefficient is +/- 1.0.
- Multicollinearity among independent variables will result in less reliable statistical inferences.
- It is better to use independent variables that are not correlated or repetitive when building multiple regression models that use two or more variables.
- The existence of multicollinearity in a data set can lead to less reliable results due to larger standard errors.
- Sys_bp, dia_bp, BMI, glucose, heartrate, age, total_chol poses to be highly with dependent variabels which is not good to perform models on top of it.

Encoding of categorical features

Machine learning models can only work with numerical values. For this reason, it is necessary to transform the categorical values of the relevant features into numerical ones. This process is called feature encoding.

Feature Scaling

Machine learning is like making a mixed fruit juice. If we want to get the best-mixed juice, we need to mix all fruit not by their size but based on their right proportion. We just need to remember apple and strawberry are not the same unless we make them similar in some context to compare their attribute. Similarly, in many machine learning algorithms, to bring all features in the same standing, we need to do scaling so that one significant number doesn't impact the model just because of their large magnitude.

Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

So I scaled the input features by using StandardScaler.

The Metric Trap:

One of the major issues when dealing with unbalanced datasets relates to the metrics used to evaluate our model. Using simpler metrics like accuracy score can be misleading. In a dataset with highly unbalanced classes, the classifier will always "predict" the most common class without performing any analysis of the features and it will have a high accuracy rate, obviously not the correct one.

METRICS:

- **Accuracy score**: which is the ratio of the number of correct predictions tothe total number of input samples. It measures the tendency of an algorithm to classify data correctly.

- **Precision**: Precision is the ratio between the True Positives and all the Positives. For our problem statement, that would be themeasure of patients that we correctly identify having a heart disease out of all the patients actually having it.

$$Precision = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Positive(FP)}$$

- **Recall**: The recall is the measure of our model correctly identifying True Positives. Thus, for all the patients who actually have heart disease, recall tells us how many we correctly identified as having a heart disease.

$$Recall = \frac{True\ Positive(TP)}{True\ Positive(TP) + False\ Negative(FN)}$$

- **F1 Score**: There are also a lot of situations where both precision and recall are equally important. For example, for our model, if the doctor informs us that the patients who were incorrectly classified as suffering from heart disease are equally important since they could be indicative of some other ailment,then we would aim for not only a high recall but a high precision as well. In such cases, we use something called F1-score. F1-score is the Harmonic mean of the Precision and Recall:

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Class Imbalance Issue :

In this problem we have a dataset of patients where we have to find out whether the given features or symptom a person has, he/she has a cardiovascular disease in future. But here's the catch… therisk rate is relatively rare, only 15% of the people have this disease.

Random Over-Sampling :

Oversampling can be defined as adding more copies to the minority class. Oversampling can be a good choice when you don't have a ton of data to work with. A con to consider when under-sampling is that it can cause overfitting and poor generalization to your test set.

SMOTE-TOMEK:

This method combines the SMOTE ability to generate synthetic data for minority class and Tomek Links ability to remove the data that are identified as Tomek links fromthe majority class, that is, samples of data from the majority class that is closest with the minority class data. We only oversampled the train data, test data remain untouched form making synthetic duplicates. The reason being since the oversampling technique will introduce data points near current

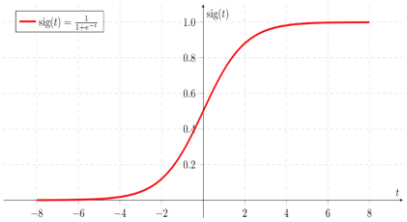
data points belonging to same class which may not accuratelydepict your test data. If you balance the Validation set (test data), your model may work well (may get better score in Val) butin the future after deploying, it may not work better so while training, validate with imbalance data only.

Confusion Matrix:

A Confusion matrix is an N x N matrix usedfor evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model.

Logistic Regression

Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0or 1, true or False, etc. but instead of givingthe exact value as 0 and 1, it gives the probabilistic values which lie between 0and 1.



Decision Tree

A decision tree is a type of supervised machine learning used to categorize or make predictions based on how a previous set of questions were answered. The model is a form of supervised learning, meaning that the model is trained and tested on a set of data that contains the desired categorization.

The decision tree may not always provide a clear-cut answer or decision. Instead, it may present options so the data scientist can make an informed decision on their own. Decision trees imitate human thinking, so it's generally easy for data scientists to understand and interpret the results.

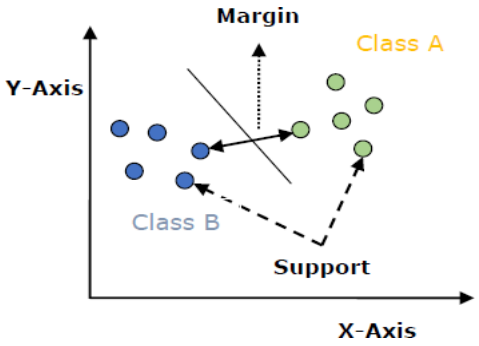
Random Forest

The Random Forest classifier creates a set of decision trees from a randomly selected subset of the training set. It is basically a setof decision trees (DT) from a randomly selected subset of the training set and then it collects the votes from different decisiontrees to decide the final prediction.

In Laymen's term, the training set is given as: [X1, X2, X3, X4] with corresponding labels as [L1, L2, L3, L4], random forest maycreate three decision trees taking input of a subset of it.

Support Vector Machine

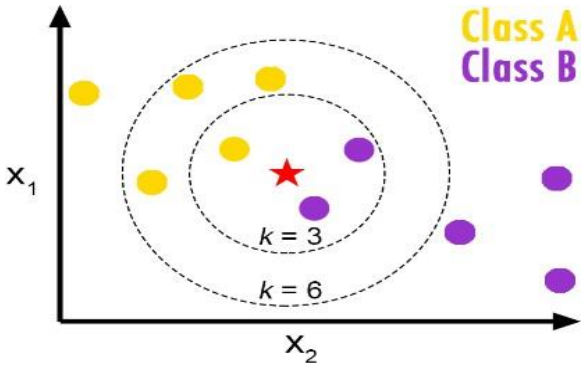
An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner bySVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).



K-Nearest Neighbour

- Instance-based learning: Here we donot learn weights from training data to predict output (as in model-based algorithms) but use entire training instances to predict output for unseen data.
- Lazy Learning: Model is not learned using training data prior and the learning process is postponed to a time when prediction is requested on the new instance.
- Non -Parametric: In KNN, there is no predefined form of the mapping function.

The distance between the two data points is calculated by the following methods Euclidean distance**, Hamming** distance, Manhattan distance, Minkowski distance. Euclidean is most popular amongst all.



Naïve Bayes

It is a probabilistic classifier, which means it predicts on the basis of the probability ofan object. It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence ofother features. Such as if the fruit is identified on the bases of colour, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence eachfeature individually contributes to identify that it is an apple without depending on each other. It is called Bayes because itdepends on the principle of Baye'stheorem.

The formula for Baye's theorem is:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis istrue

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Conclusion:

By fitting the data into various classification models and evaluating with test data, we can conclude that,

For our problem statement, we have two cases

- **Case 1 :** if a person has a disease but the model shows the person has no risk. For this case **recall** is the best evaluation metric
- **Case 2 :** if a person has no risk of disease but the models predicts the person has a risk. For this case, **precision** is the best evaluation metric
- When both of the above cases are important, we can use **F1_score** evaluation metric
- If **Recall** is important, i.e., if we want to focus on case 1, then **Logistic Regression** and **KNN** models are preferrable
- If **Precision** is important i.e., if we want to focus on case 2, then **Naive Bayes** model is preferrable
- If **F_Score** is important, i.e., if we want to focus on both case 1 and case 2, then **Logistic Regression** model is preferrable.